Chapter 2

THEORETICAL BACKGROUND AND LITERATURE REVIEW OF COMPRESSION & ENCRYPTION TECHNIQUES

2.1 PRELIMINARIES OVERVIEW

This section discussed the following basic requirement for completion of this work

2.1.1 File format and size of DNA sequence

The text file have within a group of four symbols (c,g,t & a) coming one after another & ending with blank space mean the end of text file. The text file is the fundamental part in compression-encryption and decryption-decompression process. The output place for keeping records also text file, has in it the information given of both unmatched four base pair and a coded value of American Standard Code for Information Interchange (ASCII) character. The coded value is written into the output file. The input-output file size is measured by byte and depends on the number of bases present in the input output text.

2.1.2: DNA sequence substring formation process

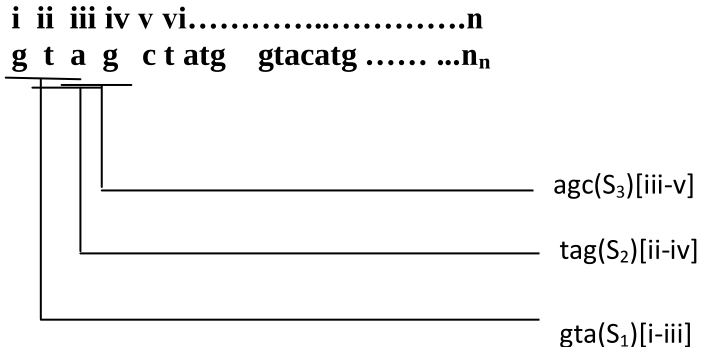
i ii iii iv v vi…………..…………n
g t a g c t atg gtacatg …… ...n_n
agc(S_3)[iii-v]
tag(S_2)[ii-iv]
gta(S_1)[i-iii]

Fig. 2.1 substring formation process

From the pictorial representation of fig. 2.1 shows that for sth substring Ss.
"s" is the substring starting position and l= (s-l) + l, is substring ending position, where l is the length of the substring.

2.1.3 Mathematical formulation

The finite string over the symbol (c,t,g & a) is defined by the alphabets s, e, |s| define the string length of s, and the symbol number in s.s_f is the fth character of s. $s_{f:l}$ is the substring of s from position f to position l. The first symbol of s is s_f. Thus $s=s_{1:|s|-1}$, where $s_{f:l}$ defined the substring of original sequences and |e| defined the string length of e, the symbol number in e.e_f is the fth symbol of e.$e_{f:l}$ is the another substring of e from position f to position l. The first symbol of e is e_1. Thus $e=e_{1:|s|-1}$. $s_{f:l}$ match with $e_{f:l}$. The length of substrings is defined by s-e which is minimum difference. The $e_{f:l}$ represents the palindrome, repeat, genetic palindrome & reverse substring. If exact match is found then, $s_{f:l}= e_{f:l}$ and count exact maximum repeated of $s_{f:l}$. In case of string is empty, define by Є =0.

The most common string searching problem is to find all occurrences of a string $P=p_1p_2p_3…p_m$ inside a large text find $T=t_1t_2t_3….t_n$, where m is the length of the DNA sequence P, n is the number of subsequence in exact match and T is the library file. We assume that the string and the DNA sequence of 4 nos of characters from a finite character set ∑ . The pattern of DNA are not exact, we may not remember the exact pattern in all places in the long string. The approximate string matching problem is to find all substrings in T that are equal to P under some measure of equalness.

Consider a string (DNA sequence as a string) s=atggtagtaatgtacatgcatg……..n, where n is defined by the file size, the sub-sequence size depends on user requirement. First break the string in sub-sequences as $s^{/}=s_1s_2s_3……..s_n$ where s_n is defined by the sub-sequence number. Then first s1 sub-sequence matching with the whole string s, then s_2 matching with the whole string s and so on up to s_n sub-string matching with the string s, after that find which sub-string pattern matching in maximum places on the string s, this process is going on until and unless the overlapping does not occur.

2.1.4 Algorithm evaluation

- **Accuracy**

In order to store a DNA sequence first required perfection. If change a single base DNA order in structure, addition & subtraction would outcome in very great impact of phenotype as seen in the Sicklemia. It is unable to put up with any error currently in existing either in compression- encryption and decryption-decompression. Mathematically not extablished but proved by text record mapping that all techniques in this algorithm are errorless, since each base order uncommonly correlate to an ASCII code.

- **Efficiency**

This technique can compress source files on the basis of the substring of length into 1 characters for all the DNA part, and the output contains less character than input DNA sequence characters

- **Space occupation**

These techniques operate on very small constant memory space, requiring minimum of 512MB main memory. In these techniques where input process reads character one by one and place them into the output file immediately.

2.1.5 Working principal

1. File type: All DNA sequences are in text format, the file extension is dot txt.

2. Subsequences are auto generated by breaking the query sequence into words

3. The DNA sequence is encoded by some edit process

Our approximate matching process operates by two standard edit operation as

1) First is replace-process is defined by (C, P, S.) where C is the replaceable symbol at position P by the symbol S.

2) Second is insert - process is defined as (L,P,S) where symbol S insert into position P of the length L.

2.1.6 Hardware and software specification

Since the program is written originally in the C language based on Turbo C Compiler and operating platform is Windows XP. This program is also run on other computer required little bit changes (being dependent on the operating system and translator used). The list of programs run on the IBM personal knowledge processing machine have need of 512K, without adding of hardware except for thin, flat, round plate drivers and printer.

2.1.7 Evaluation parameter

The performance of this algorithm is analyzed by the following parameters as below

- The **compression ratio** is defined by; 1− (|Output(O)|/2| input(*I*)|)

- **The defination of compression rate** is (|output(O)|/|input(I) *I*|),

- **Saving Percentage**: is calculated as = (I-O/I)*100%

Where |O| is the output file size measured in bits and |I| is the input file size measured by bases in the sequence.

●The improvement Y over X is defined as (X-Y)/X *100, Where Y is the developed algorithm result and X is the existing standard algorithm result.

● The average speed of compression and decompression is measured by millisecond per input byte.

● **The Compression gain**:

Gain = original size - encoded size, and rate

$$\text{Average compression gain} = \frac{\text{Gain}}{\text{Original size}} \text{ X100}$$

$$\text{Rate} = 1 - \frac{\text{Gain}}{\text{Original size}}$$

● **Complexity in space**

The amount of space required by an algorithm to complete its execution is called as complexity in space.

● **Complexity in time**

The total time required to complete an algorithm to run

● **Encryption ratio(ER):** This is defined by encrypted part divided by whole data size.

● **Compression friendliness (CF)**: If an encryption process has no impact on data compression accuracy, then this algorithm is considered compression friendly.

●The **effect on actual text**: is measured by Levenshtein Distance divided by whole file size.

Levenshtein Distance is calculated by dissimilarity between two strings inserting or deletion of character.

● **Time in encryption**

The time in encryption is defined by, how much time is required to convert plain text to cipher text, process of encryption depends upon plain text block size, mode and key size. The time in encryption is measured in 10^{-3} second of our experiment. The performance of the algorithm has great influence on encryption time.

- **Time in decryption**

The time required to get back cipher text to plaintext is called time in decryption. The process of decryption required less time than time in encryption to make system sensitive and quickly moving. Time in decryption is measured in 10^{-3} second in our experiment. The performance of the algorithm has great influence on decryption time.

- **Rate of encryption** is defined by file size divided by encryption time

- **Avalanche effect** is defined by hamming distance divided by file size. Hamming distance is measured by similarity between two strings.

- **DNA sequence entropy**

The important property is randomness of compression-encryption method and hard to track by a hacker. The information Randomness is measured by the entropy and its uncertainty. In the field of information security required high randomness after encryption as a result, there is less or no dependency between key and cipher text. If the information contained more randomness, the complex situation occurs in between key and cipher text, it is difficult to guess by an attacker. Entropy reflects performance of compression-encryption algorithms. The entropy is calculated by using Shannon's formula.

$$H(X) = -\sum_{i=0}^{N-1} p_i \log_2 p_i$$

Where p_i is the probability of a given symbol.

- **Impacts cost**

The encrypted bit after encoding will be transmitted over unreliable networks, this is called required bandwidth transmission. If fewer number of bits are encoded, it will consume less storage and lesser bandwidth. This is called impacts cost.

- The **Encryption throughput** (Byte/See) is defined by = ∑Encryption file size / ∑ Encryption time.

- The **Decryption throughput** (Byte/See) is defined by = ∑Decryption files size / ∑ decryption time.

2.1.8 Decompression technique in client side

Most of the compression algorithm aim at high compression rate with decompression time is less is mostly liked for effective order of facts with quicker data transmission. Our aim is to lessen compression & decompression time and provide high information safety. DNA search engines currently in existence do not make use of DNA order compression algorithms & encryption for high safety for decryption & decompression at client end i.e where an encrypted compressed DNA order is changed back into starting form & decompression at the person for whom one does work is furnished for help of quicker sending & information safety because the purpose of many DNA orders now a days is to attain high rate during compression or pattern unveiling than decryption decompression at client end. Their time of decompression is lengthy in comparison to necessary information security.

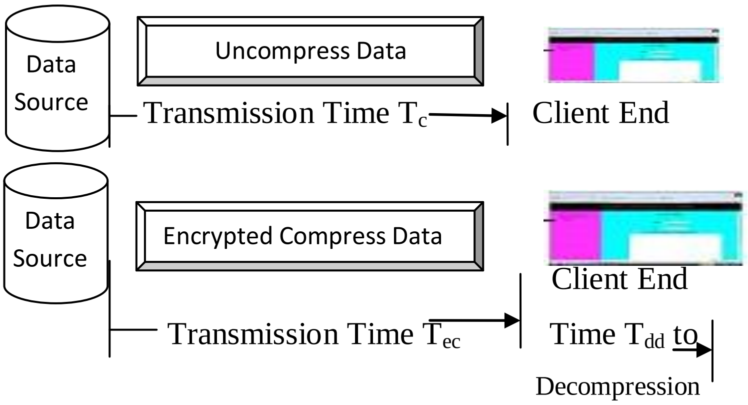

Fig.2.2 Show the client site encryption & decryption process
Efficiency is achieved when, $T_{ec} + T_{dd} < T_c$

In order to make a comparison of the overall operation, further studies were conducted involving transmission of actual order file of changing sizes to calculated time(T_c) required for sending from the starting point to the place where on it is going. Then the files are compressed utilizing both compression-encryption and decrypt decompression algorithm. T is the total time formed as time for transmission (T_{ec}) of the encrypted compressed text with decrypt decompression time(T_{dd}) at client end is calculated by both these processes. The relation of time for decompression and time for transmission of the original uncompress sequence (T_{dd}/T_c) gets changed to other form with the file size increased. The efficiency is achieved if the encrypted compress file transmission time plus client side decrypt decompression time is less than the calculated time in between source to destination transmission time. Then we can get the efficiency as shown in fig. 2.2

2.2 THEORETICAL BACKGROUND

2.2.1 Introduction of Genomics

Genomics in the field of interdisciplinary science, genomes mainly aim function, structure, mapping, evolution, mapping, and genome editing. A group of DNA of an organism's is called genomes, in addition to its genes. A DNA produces protein with the help of enzyme and molecules as messenger. Genomics is analysis by high throughput DNA sequencing and Bioinformatics analyze the entire genomes function and structure.

The information theory provided the compression related theoretical background by Claude Shannon [9], and published fundamental papers on the idea of compression in the year 1940 and 1950. Life is organized and based on structured [34]. The organism having single cell contain large amount of information into a DNA molecule. This huge amount of data required efficient storage, to remove the redundancy of the data.

Bioinformatics mostly used comparative analysis and computational methods of genomic data from 1980s onward. Bioinformatics is used to solve different biological problem [35] by using and developing algorithms and computational tools. It is useful to solve the problem such as compression of huge DNA sequences, unknown DNA similarity searching, protein similarity searching, protein function prediction and 3D protein structure prediction. It is helpful to researchers for more precise experiments or design better to solve the biological problem. The existing biological databank is used in Bioinformatics for analyzing raw data derive from various experiments. Thus, it is seen that the biological databank is very important in Bioinformatics.

In recent times the volume of biological data has increased tremendously as such there is demand for Bioinformatics tools for compression [36] and proteomic and genomic data analysis. One of the Bioinformatics tools is web based platforms used for biological data analysis in the field of bioscience.

The genetic information is transferred from parents to offsprings by DNA. Some part of DNA translates protein, which is important for the structure and function of cellular life. At present there are three practical problems in biological database- storage, security and comprehension. At present in text format the DNA and protein sequence is stored, in text format each base required eight bits for storing, required huge space. After encoding the DNA and protein sequence size is reduced to a half or a quarter of its current size.

2.2.2 Growth of DNA sequence

The graphical representation shows that the human genome sequence versus annual sequence capacity is exponentially increased. The capacity of DNA is defined by Tera basepairs(Tbp), Peta basepair (Pbp) and zetta basepairs(Zbps). In the year 2015, the size of genomes is change noticeably. Many genome project submit the huge genome sequence such as human genome [37], cancer genome atlas[38] etc.

2.2.3 What is DNA

The full form of DNA is deoxyribonucleic acid. All organism information related to build and maintain is stored in DNA, which is also a complex molecule. DNA is the genetic code that determines all the characteristics of a living thing. DNA has four bases. They are adenine (A), guanine (G), cytosine (C) and thymine (T). The DNA provides all genetic information. The DNA is used in our database as a biological data and information store for a long term. It is often compared to a set of blueprints, like a recipe or a code. The DNA has some instruction for producing RNA molecules and proteins. The gene, DNA segments carries the genetic information .

2.2.4 History of DNA

James Watson, an American biologist and Francis Crick a British physicist first discovered the DNA in the year of 1950s. They presented a model of the DNA double helix. This model is also known as Watson and Crick models . They won the Nobel Prize in 1962 for their model. Rosalind Franklin discovered a powerful technique for determining the structure of molecules and it is also known as X-ray crystallography. Levene discovered the four bases: adenine (A), thymine (T), guanine (G), and cytosine (C) in 1919

2.2.5 Properties of DNA

Molecules are also called nucleotides which are made up of DNA. Every nucleotide contains the nitrogen base, phosphate and sugar group. The types of nitrogen bases are four. They are: thymine (T), adenine (A), cytosine (C) and guanine (G).The nitrogen bases are forming A-T and G-C base pairs. A-T contains two hydrogen bonds base pairs and G-C contains three hydrogen bonds base pairs. 10 ångströms (1.0 Nanometre) is the radius of the pair of chains and one nucleotide unit measured 3.3 Å (0.33 nm) long.

2.2.6 Define compression

Lossless and Lossy compression are the two types of data compression technique which are used to compress the data.

Lossless Compression

In this process of compression, the original sequence will be getting back after decompression identical with the compressed data. In Lossless Compression technique no data will be lost in case of encryption and decryption of any message.

Lossy compression

It is a one type of Data compression technique. In Lossy Compression technique some data will be lost in case of encryption and decryption of any message.

2.2.7 Text- Vs DNA compression

Gzip or zip is used to compress text file as well as text files containing DNA.

8 bits are used for each character in a text file. In case of DNA, two bits/ base A = 01; C = 10; G = 01; and T = 00 is mostly used for encoding. Using this process, all files contain DNA base have increased the file size without compression.

2.2.8 Information Theory: Biological Information

The DNA molecule is defined by symbol having to the alphabet {T,A, G,C}.Here each symbol represents a nucleotide and they have equal probability and all sequence symbols are independent. The basic principal of information theory is used to assert that 2 bits/symbol is required for coding sequence of a computer and Maximum Entropy = $\log_2$ (n) | n =the number of possible states. For an example given that the addition of start and stop codons, three in total and the number of codons is 61 that give content to exons, so the maximum of entropy per nucleotide for this reality is log2 (61 / 3) ≈ 1.977 bits.

2.2.9 Entropy coding in Genomic Sequences

It was published in the seminal book[39] in the year 1972 the DNA sequence entropy analysis is introduced. It is the relationship between biological information and information theory. It is also applicability of the concepts of entropy in the analysis of DNA sequences. The entropy [40], redundancy, complexity and compression are same in concepts. Studying these concepts is very challenging in DNA sequences, there have about average 99% of unpredictability and it is not clearly understood practically nothing of the complexity of its language. Entropy is very close to 2 bits per base[41] in the case of Bacteria present levels.

2.2.10 Pattern Discovery

MDL principle is used to discover the biological patterns in Bioinformatics[42]. Data compression techniques and the associated is the aim of this section .

2.2.11 Data Set

GenBank is a very big archival database where sequencing labs submit their data. Databases are guided by the INSDC. The recommended GenBank is the Microarray database Arry Express[43], the DNA Data Bank of Japan(DDBJ)[44], the European Bioinformatics Institute EMBL database[45] and the National Centre for Biotechnology Information(NCBI)[46] etc.

The standard benchmark data are used by the DNA compression algorithm. The data comes from several sources, including the complete genomes [47] 2 CHLOROPLASTS (CHMPXX and CHNTXX “in addition called MPOCPCG”). Five orders from humans (HUMHBB , HUMDYSTROP, HUMHPRTB, HUMGHCSA and HUMHDABCD, 2 causes of diseases (HEHCMVCG and VACCG and “in addition called HS5HCMVCG”) and finally 2 mitochondria (PANMTPACGA and MPOMTCG, “in addition called MIPACGA”) .

2.2.12 Defining repetition

The recurrence of a pattern in a sequence is called repetitions. A DNA sequence has so many repeats. In four ways a DNA pattern is described as indirect, direct, reverse complement and complement. In a direct repeat a pattern is in the same stand of a same nucleotide in many places such as ACCG repeats as ACCG in a DNA sequence. In indirect repeat such as ACCG is GCCA is repeated in many places of DNA sequence. A complement of ACCG is TGGC repeated in the same sequence in many places. Here complement A is replaced by T and G is replaced by C and reverse complement repeat of ACCG is CGGT is repeated in the same sequence in many places[48].

If a sequence ATAT is reversed it will be ATAT i,e the character position left to right and right to left are same. This special recurrence is called palindrome. It is shown in consecutive word or phrase.

2.2.13 DNA repetitions in biological classes

Many biological literature shows that the DNA sequence has so many repetitions in classification schemas. Different techniques measured for each schema classified repetitive DNA characteristics. Satellites, minisatellites and microsatellites, proretroviral transposons and retroposons are the four broad classesof these systems.

2.2.14 String matching

The string matching classical technique of finding strings that match a pattern approximately or rather than exactly in science and technology. Naive is an example of string matching technique. String matching is a technique used to match the strings.

2.2.15 Exact repeat string searches

The substring repetition in a long string is not a easy task.

n= length of the given string and s = substring, where r an exact repeat can be defined as $s^c s^/$ where c consecutive occurrences of s are followed by a prefix of s,$s^/$.

2.2.16 String compression

Huffman coding[49], arithmetic coding[50] and Lempel-Ziv coding[51] algorithms are used to compress the string. By the process of substring recurring, the string is compressed and reduce the space and time of compression. String compression is also used to compress big DNA sequences to a small DNA sequence

2.3 LITERATURE REVIEW OF EARLIER WORK ON COMPRESSION

Shannon information theory and science about living things have a long time disputable relationship [52]. The theory of information technology has been successfully utilized for many year for order observation and in specific to count the amount of ‘divergence from randomness’ of DNA order [53].

The typical application of the theory of information technology is textual data compression. It is deeply interconnected to statistics, classification and different ideas related to order complexity [54]. In the past ten year the successful outcome of computational biology is sequence compression.

Nowadays, so many compression algorithms are offered on the basis of the special structures of DNA order.

Lossless data compression techniques are used to compress files or data into a smaller form. It is often used to pack up software before it is sent over the Internet or it is downloaded from website to cut down the amount of time and bandwidth required to transmit the data. The lossless data compression techniques has the constraint when data is uncompressed, it must be same to the actual data that was compressed.

2.3.1 Encoding based on entropy

It is a compression algorithm. The number of times each alphanumeric character is repeated in the given text is checked by this algorithm. After that it replaced by a unique character. The length of the given input file varies with the frequency of the symbols.

Huffman coding: Huffman coding is an entropy based technique. In this technique compress the average code length of symbols into alphabet. This algorithm developed by David A. Huffman, is published in the 1952 paper “A Method for the construction of Minimum Redundancy Codes”. In information theory and Computer Science, a Huffman code is used for lossless data compression. Statistical Code represents fixed length data blocks with variable-length code words. One type of statistical code is Huffman coding.

Arithmetic Coding: The arithmetic coding technique is developed by IBM. This algorithm is a process of encoding based on entropy, which is applicable on lossy and lossless data compression. The highly repeated symbol is replaced by a small number of bits encoding process than rarely seen symbols. It has some advantage techniques like Huffman Coding, but it has also some drawbacks.

2.3.2 Dictionary based encoding

Dictionary based encoding developed by Jacob Ziv and Abraham Lempel in 1977.It is the first popular universal compression algorithm for data when no prior knowledge of the source was available. The LZ77 (and variants) is mainly used in many popular compression programs such as ZIP and GZIP. It has a circular buffer called the “Sliding window” which holds the last N bytes of data processed. LZ78 is used for more definite dictionary structure.

2.3.3 Substitution Based Methods

Substitution base methods are used for converting a plaintext into cipher text. In the year 1993 Grumbach and Tahi[55] developed the first special purpose DNA compression . This is found on the LZ algorithm.

2.3.4 Substitution and Statistical Based Methods

Substitution and statistical methods are a several DNA compression algorithms. The Off-line algorithm was developed by Apostolico and Lonardi[56]. It is used in repeated regions of compression.

CTW + LZ algorithm was developed by Matsumoto et al. and Tabus et aldevelopeda sophisticated DNA sequence compression.It was based on NML.

This work was improved into GeNML in the year of 2005. Mishra et al. developeda DNASC or DNA Sequence Compressoris Publishedin 2010.

2.3.5 Compressed Pattern Matching

Searching for pattern in compressed data with little or no decompression in computer science and information technology by Compressed Pattern Matching or CPM. It is easy to search a compressed string quicker than an uncompressed string by us .It also requires less space. A hybrid compression technique is developed by G. Navarro and M. Raffinothas. It allows fast searching as LZ78 and at the same time maintains many of the features of LZ77.

2.4 LITERATURE REVIEW OF EARLIER WORK ON ENCRYPTION

The cryptography was since from World war-II. The term cryptography is derived from a Greek word kryptos which means “secret writing”. In order to protect information from unauthorized user an effective tool is used called Cryptography. It is used for computer security in many aspects.

Various components involved in cryptography are Plain text, Cipher text, Encryption, Decryption, Ciphers, Key.

Cryptanalysis is the science and art of breaking the encrypted codes that are created by applying some cryptography algorithm. The person who performs Cryptanalysis is known as Cryptanalyst.

2.4.1 Historical Development of Ciphers

In the Second World War, electromechanical and mechanical cipher instrument came into being. Kahn [106] developed classical ciphers and instrument within 1967. The classical ciphers mainly follow the Shannon's paper published in the year 1949[43].The telegraph encryption was developed by Vernam cipher [57]. Proposed matrix cipher of Hagelin M-209 Hill [58], technical details was written by Berker & Piper[59] and provide practical method for poly-alphabetic substitution.

The rotor machines instructive overview is given by Diffie and Hellman[60] used in high level system.

2.4.2 Data Encryption Standard (DES)

DES is a symmetric key cipher. It was developed by IBM. The original specification of DES is the 1977 U.S. Standa Feistel[61] Cipher implementation version is Data Encryption Standard or DES. 16 round Feistel structureis checked by DES it has also 64-bit block size.

2.4.3 AES (Advanced Encryption Standard)

In the year 1977, the U.S. government has published Advanced Encryption Standard or AES. Same key is used for encryption and decryption the message in this algorithm .AES comprises three block ciphers, AES-128, AES-192 and AES-256.

2.4.4 Asymmetric key encryption

Asymmetric key encryption algorithm, also known as public/private key pairs. Public/private key pairs are used for asymmetric encryption. In like in size key process of changing knowledge into a secret form of algorithm, it is necessary to make distribution of the key before the process of changing knowledge into a secret form and process of changing knowledge back into starting form because both purposes used the same key.

2.4.4.1 Rivest, Shamir & Adelman

Rivest, Shamir &Adelman (RSA)[62] is a relatively slow algorithm. Rivest, Shamir &Adelman or RSA algorithm is a public key encryption technique and is considered as the most secure way of encryption. It was invented by Rivest, Shamir and Adleman in year 1978 and hence name RSA algorithm. The private & public key are produced by RSA algorithm.

2.4.5 Public-Key Encryption

Public-key cryptography is an encryption technique that uses a pair of public and private key algorithm for secure data communication. The public key is used for encryption of the message and decrypt is done by the private key.

2.4.6 Digital Signature

The authentication & message integrity is checked by digital signature.